強化學習 HW2 0751231

Problem 1.

(ii)

$$L_{\pi_{\theta_1}}(\pi_\theta) = \eta(\pi_{\theta_1}) + \sum_s d_\mu^{\pi_{\theta_1}}(s) \cdot \sum_a \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a)$$

$$\Rightarrow \nabla_\theta L_{\pi_{\theta_1}}(\pi_\theta) = \sum_s d_\mu^{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a)$$

又, $\eta(\pi_\theta) = \eta(\pi_{\theta_1}) + \sum_s d_\mu^{\pi_\theta}(s) \cdot \sum_a \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a)$

$$\Rightarrow \nabla_\theta \eta(\pi_\theta) = \sum_s \nabla_\theta \left(d_\mu^{\pi_\theta}(s) \cdot \sum_a \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a) \right)$$

$$= \sum_s \left[\left(\nabla_\theta d_\mu^{\pi_\theta}(s) \right) \cdot \left(\sum_a \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a) \right) + d_\mu^{\pi_\theta}(s) \cdot \left(\sum_a \nabla_\theta \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a) \right) \right]$$

$$\Rightarrow \nabla_\theta \eta(\pi_\theta)\Big|_{\theta=\theta_1} = \sum_s \left[\left(\nabla_\theta d_\mu^{\pi_\theta}(s) \right) \cdot \left(\sum_a \pi_{\theta_1}(a|s) \cdot A^{\pi_{\theta_1}}(s,a) \right) + d_\mu^{\pi_{\theta_1}}(s) \cdot \left(\sum_a \nabla_\theta \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a) \right) \right]$$

$$= \sum_s d_\mu^{\pi_{\theta_1}}(s) \cdot \sum_a \nabla_\theta \pi_\theta(a|s) \cdot A^{\pi_{\theta_1}}(s,a)$$

$$= \nabla_\theta L_{\pi_{\theta_1}}(\pi_\theta)$$

(i)

$$L_{\pi_{\theta_1}}(\pi_{\theta_1}) = \eta(\pi_{\theta_1}) + \sum_s d_\mu^{\pi_{\theta_1}}(s) \sum_a \pi_{\theta_1}(a|s) A^{\pi_{\theta_1}}(s,a)$$

$$\approx \eta(\pi_{\theta_1}) + \left(\eta(\pi_{\theta_1}) - \eta(\pi_{\theta_1}) \right) = \eta(\pi_{\theta_1})$$

Problem 2.

$$D(\lambda) := \min_{\theta \in R^d} \left\{ -\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right)^T (\theta-\theta_k) + \lambda\left(\frac{1}{2}(\theta-\theta_k)^T H (\theta-\theta_k) - \delta \right) \right\},$$

H is a positive definite matrix and hence $L(\theta,\lambda)$ is strictly convex.

$$\nabla_\theta L(\theta,\lambda) = -\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right) + \lambda H(\theta-\theta_k)$$

Since $L(\theta,\lambda)$ is strictly convex, then a point θ^*, λ^* the global minimum if and only if $\nabla_\theta L(\theta,\lambda)\Big|_{\theta=\theta^*,\lambda=\lambda^*} = 0$

$$\Rightarrow \nabla_\theta L(\theta^*,\lambda^*) = -\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right) + \lambda^* H(\theta^*-\theta_k) = 0$$

$$\Leftrightarrow \theta^*-\theta_k = \frac{1}{\lambda^*} H^{-1}\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right)$$

Hence, $D(\lambda) = L(\theta^*,\lambda) = \frac{1}{\lambda^*}\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right)^T H^{-1}\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right) +$

$$+ \lambda^*\left(\frac{1}{2\lambda^{*2}}\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right)^T \cdot H^{-1} \cdot H \cdot H^{-1} \cdot \left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right) - \delta \right)$$

$$= \frac{-1}{2\lambda^*}\left(\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right)^T H^{-1}\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right) \right) - \lambda^*\delta \quad \#$$

$$\lambda^* = \sqrt{\frac{\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right)^T H^{-1}\left(\nabla_\theta L_{\theta_k}(\theta)\Big|_{\theta=\theta_k} \right)}{2\delta}} \qquad (b)\ \alpha = \frac{1}{\lambda^*}$$

#Lambda*是使用算幾不等式求得

Problem 3 :

(a)

Hyperparameters :

Learning rate : 0.005

step_size=100

gamma=0.9

Loss function :

Policy loss = log_prob(action) * advantage

Value loss = L1_smooth_loss

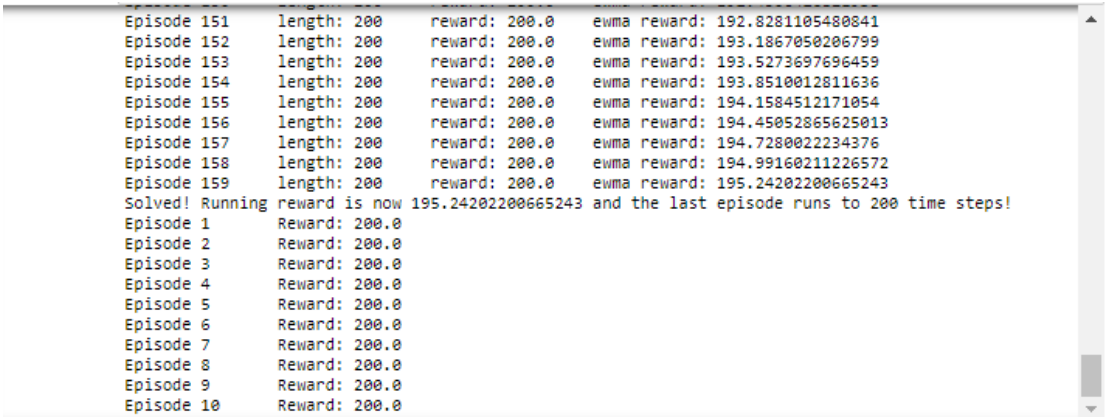
NN architecture :

Policy : Input(4) , FC(128) , relu() , FC(128) , relu(), FC(2) ,softmax()

Value : Input(4) , FC(128) , relu() , FC(256) , relu(), FC(1)

Implement result :

Achieve reward threshold at episode 159



(b)

With MC:

Hyperparameters :

Learning rate : 0.01

step_size=100

gamma=0.9

Loss function :

Policy loss = log_prob(action) * advantage

Value loss = L1_smooth_loss

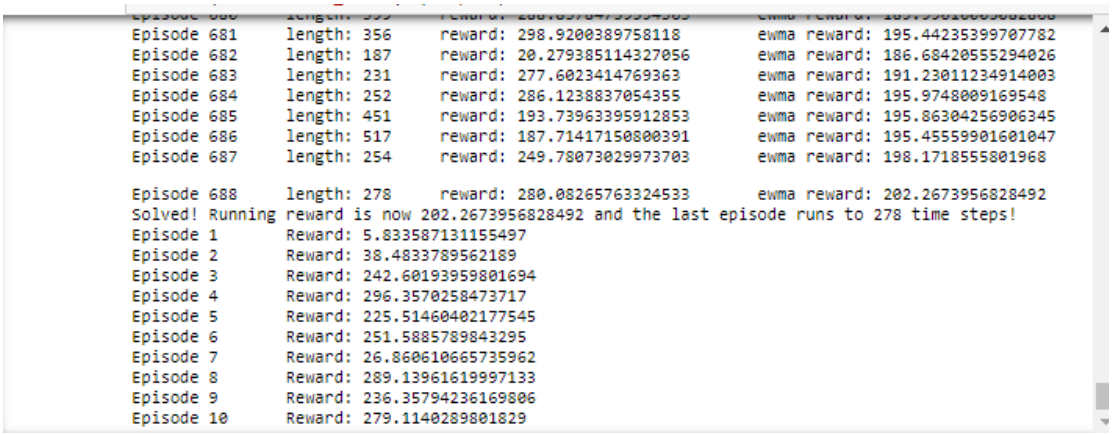
NN architecture :

Actor : Input(8) , FC(128) , relu(), FC(2) ,softmax()

Critic : Input(8) , FC(128) , relu() , FC(1)

Implement result :

Achieve reward threshold at episode 687



With TD(0) bootstrapping:

I tried to tune the hyperparameters many times and tried to split actor and critic. But always did’t get a acceptable performance.